

Chapter 6

Data Prediction and Suggestive Measures

6.1 Introduction

The process of generating new data points based on the relationships or the trails present in the existing data points is referred to as data prediction. The models that are used to perform prediction are known as prediction models. These models can be either statistical, non-statistical or hybrid models based on the context. Non-statistical models are based on the identification of patterns that are present in the data, whereas statistical models are based on the formulation of heuristics between the participating variables based on probability. Statistical models that are highly interpretable, simple to apply, and applicable to small datasets; nevertheless, they are quite confined. On the other hand, models that are non-statistical are highly complicated and are applicable to bigger datasets; however, they are incredibly flexible. Linear regression and autoregressive integrated moving average (ARIMA) are two examples of statistical models. Random Forest and Agent-Based Modeling (ABM) are two examples of models that are not statistical in nature [226].

6.2 Some of the Statistical and Non-Statistical Prediction Models

6.2.1 Statistical Models

Statistical models aim at predicting/ forecasting unseen values for future by estimating the possible relationships between values in the past (i.e. it establishes relationship between dependent variable (response) and the independent variable (s) (Predictor) for making prediction.

a) Linear Models

Linear models assume linear relationship between dependent and independent variable(s).

- The linear regression model is a statistical model that is used to predict continuous data points by assuming a linear relationship between a dependent variable and one or more independent variables. Depending on the number of independent variables, the regression model takes either of these two forms: simple (one independent variable) or multiple regression (more than one independent variable). Multiple regression, in addition, assumes multicollinearity; i.e., there should not be a high correlation between the independent variables.
- Generalized Linear Models (GLMs) are an adaptation of linear regression that applies linear regression to non-normal distributions. Some of its types are logistic regression, Poisson regression, and negative binomial regression. Logistic regression is used when the model is expected to produce binary values as an outcome. Poisson regression is used when the model is expected to produce a non-negative count. Whereas negative binomial regression is used when the data when the model is expected to produce a non-negative count when variance is greater than the mean and violate the Poisson assumption. A unique kind of statistical model called Dirichlet Regression is used to analyze multivariate outcomes in which the components represent proportions that add up to one; in other words, it can manage interdependent proportions among all the components at once while maintaining their proportional nature.

b) Time Series Models

This model considers dependencies between dependent variable and its temporal instances for the purpose of prediction.

- The AR model (autoregressive model) is a typical example of a time series model wherein in the future values are predicted based on the values attained by the same variable over a period of time. This model assumes a linear relationship between the determined value and the previous values, along with stochastic error.
- In contrast to AR models that uses the past values, the Moving Average (MA) model aims at determining dependencies in terms of the relationship between the current observations and the residual errors left from the previous observations.
- ARIMA is a very comprehensive forecasting model that combines the abilities of AR model, Integration, and Moving Average (MA). Integration is performed to make the data stationary by removing trends and seasonality.

c) Nonlinear models

Nonlinear models are used in order to identify complex relationships between dependent and independent variables which cannot take linear form or expressed as straight-line.

- Polynomial model a non-linear regression model that establishes a relationship between dependent and independent variables in a degree polynomial to allow for curve fits. Generalized Additive Models (GAMs) are highly flexible extension of Generalized Linear Model that make use of smooth functions on independent variable instead of assuming linear relations. This approach is generally preferred in situation where the relationships between dependent variable and independent variables are complex.

- A Decision Tree is a supervised machine learning that predicts by learning simple decision rules for recursively splitting data in branches or partitions.

d) Ensemble Models

Ensemble models are machine learning models that combine the capabilities of multiple base models for improving predictive performance by reducing bias, enhancing robustness by effectively handling noises and outliers, and elevating flexibility for generalization. Random forest and Gradient Boosting Machines (GBM) are some of the examples of Ensemble Models.

6.2.2 Non-Statistical Models

Non-Statistical models are models that aims at determining trends-based on the knowledge they have acquired about the expression made by the data without strict adherence to any mathematical constraints or formulations.

a) Machine Learning (ML) models

Machine Learning (ML) models are highly capable techniques that automatically detect patterns and trails in the data for making quality predictions without adhering to any explicit statistical assumptions. ML models are highly adaptive to non-linear relationships and efficiently handle complex and high-dimensional datasets. There are basically two types of ML models supervised learning models and unsupervised learning models. Some of the supervised models are Decision trees and Tree-based ensembles, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN). that split data hierarchically for predicting values, whereas k-Means and DBSCAN are examples of unsupervised models.

b) Deep Learning (DL) Models

Deep Learning (DL) Models are derivation from machine learning models that leverage upon multi-layered neural networks to identify complex patterns in data. DL models are very effective in handling high-dimensional and unstructured data for extracting features. Feedforward Neural Networks (FNNs), Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are some of the examples.

c) Rule-Based & Heuristic Models

Rule-based and heuristic models are logic-driven prediction approach that applies predefined decision rules for predicting data points rather than learning patterns from data. These models prove effective when the problem is known with a stable definition but there is insufficient data to properly train the model and there is urgency in implementation.

6.2.3 Applicability of the Data Prediction Models

The applicability of data prediction models depends primarily on the complexity of the problem, the size of the data, and its quality. Thereafter, the level of interpretability required and type of outcome expected. Statistical models work well when the problem is well-defined, interpretable, and can be modeled using heuristics; likewise, the data is structured and smaller in size. Non-statistical models are preferred in situations where there is a requirement for automatic feature extraction from the dataset that is highly unstructured, exhibiting greater dimensions, complex interactions, outliers, and omissions.

The motivation for analyzing the prediction models is to help select an appropriate model for predicting future values for the results representing various land cover classes obtained from classification of temporal satellite images presented in table 3.7, results obtained from seed-based clustering presented in table 4.2, the distribution of population participation presented in table 5.1, and the distribution of household participation presented in table 5.4.

6.3 Objectives

With respect to the research motivation, the objectives for this research initiative are:

- a) predict the changes in land cover classes obtained from classification of temporal satellite images
- b) predict the change in the density of the clusters obtained through Seed-Based clustering
- c) predict the change in distribution of population participation
- d) predict the change in distribution of household participation

6.4 Selection of datapoints

- Land cover classes obtained from classification of temporal satellite images presented in table 3.7
- Cluster densities obtained from seed-based clustering presented in table 4.2
- Distribution of population participation presented in Table 5.1
- Distribution of household participation presented in Table 5.4

6.5 Methodology pursued

The research initiatives have been strategized as per the schematic representation presented in figure 6.1,

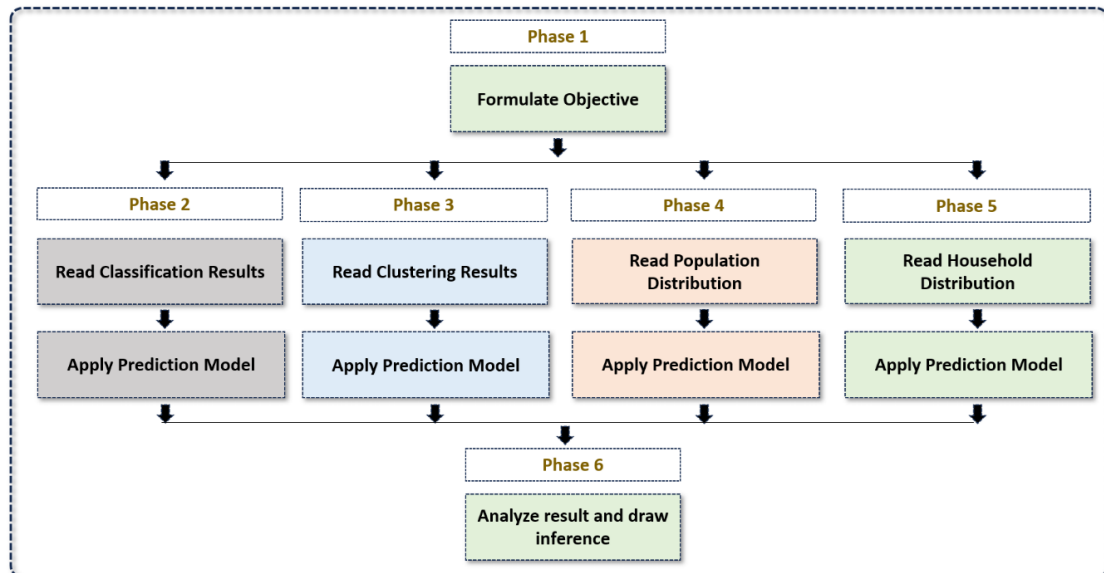


Figure 6.1: Schematic representation of the research initiative

This research initiative adheres to the following sequence of phases represented hereunder:

Phase 1: in this phase the objectives were established

Phase 2: 2.1) gather the land cover classes obtained from classification of temporal satellite images

2.2) Apply prediction model

Phase 3: 3.1) gather the cluster densities obtained as outcome of seed-based clustering
3.2) apply prediction model

Phase 4: 4.1) gather the data related to population distribution
4.2) apply prediction model

Phase 5: analyze result and draw inferences

In this research work, Dirichlet regression has been applied to the land cover classes obtained through classification of temporal satellite images; the sequence of steps has been detailed in algorithm 6.1. Linear regression has been applied for predicting change in the density of the clusters obtained through the seed-based clustering technique, distribution of population participation, and distribution of household participation; the sequence of steps has been detailed in algorithm 6.2.

Algorithm 6.1: Algorithm_Prediction_Dirichlet_Regression

Input:

- a) Land cover classes obtained from classification of temporal satellite images presented in table 3.7

Output:

- a) Model coefficients showing how each land type changes with time.
 - b) Predicted values for various Land cover classes
-

Step 1: Start

Step 2: Prepare data

- Convert pixel counts to proportions (if needed), so that values sum to 1 for each year
- Center the year variable (e.g., $Year_centered = Year - \text{mean}(Year)$)

Step 3: Define Dirichlet Regression Model (in PyMC3)

For each class “i”, define a linear model:

- Each class has its own linear predictor:

$$\mu_i(t) = \beta_{i0} + \beta_{i1} \cdot Year_{centered}$$

- Convert to positive Dirichlet parameters:

$$\alpha_i(t) = \exp(\mu_i(t))$$

-
- Model land cover proportions:

$$Y_i(t) \sim \text{Dirichlet}(\alpha_i(t))$$

- Use normal priors for the coefficients:

$$\beta_{ij} \sim N(0, 1)$$

- Fit the model using MCMC (pm.sample())

Step 4: Predict Future Proportions

- For future years (e.g., 2030), center them using the same mean.
- Use posterior beta samples to compute expected α values (Dirichlet parameters).
- Draw samples from Dirichlet distribution using those α values to simulate predictive distributions.
- Aggregate and summarize these predictions (mean, HDI).

Step 5: Fallback Approximation

If PyMC3 is unavailable:

- Fit GLM (Binomial regression) independently for each component.
- Normalize predictions to sum to 1 (approximate compositional forecast).

Step 6: Stop

Algorithm 6.1: Algorithm Prediction using Dirichlet Regression

Algorithm 6.2: Algorithm_Prediction_Linear_Regression

Input:

- a) Clustering results presented in table 4.5
 - b) Population distribution data presented in table 5.1
 - c) Household distribution data presented in 5.4
-

Output:

- a) Predicted values for density of clusters
- b) Predicted values for population distribution
- c) Predicted values for household distribution

Step 1: Start**Step 2: Read and preprocess data****Step 3: Apply Linear Regression**

- For each area (row in the table):
- Fit a linear regression model on known X (years) and y (clusters, population and household).
- Use the model to predict population values for future years.

Step 4: Store predictions

- Create a new table with area names and predicted values for 2030, 2040, and 2050

Step 5: Export results**Step 6: Stop**

Algorithm 6.2: Algorithm Prediction using Linear Regression

6.6 Results and discussion

6.6.1 Prediction of changes in the land cover classes obtained from classification of temporal satellite images

Result obtained by applying algorithm 6.1 to the land cover classes obtained from classification of temporal satellite images presented in table 3.7 is represented in table 6.1.

Table 6.1: Prediction of changes in the land cover classes by 2050

Year	Dense Forest	Vegetation	Barren Land	Urban Spaces
2000	440554	415008	141030	62391
2013	444252	411117	124103	78486
2015	439909	421416	123499	81094
2018	439045	422336	122232	82305
2023	437904	416276	122077	89661
<i>2030</i>	<i>428499</i>	<i>407181</i>	<i>111921</i>	<i>118317</i>
<i>2040</i>	<i>418906</i>	<i>398653</i>	<i>102328</i>	<i>146031</i>
2050	409313	390126	93801	172679
% increase	-7.09	-0.06	-0.33	1.77

Inference 6.1: Trend predicted in the land cover classes by 2050 (the predicted values and the trend are presented in figure 6.2 and figure 6.3)

By 2050,

- Forest cover is expected to decrease by 7.09 %
- Vegetation is expected to decrease by 0.06 %
- Barren Land is expected to decrease by 0.033 %
- Urban spawl is expected to increase by 1.77 %

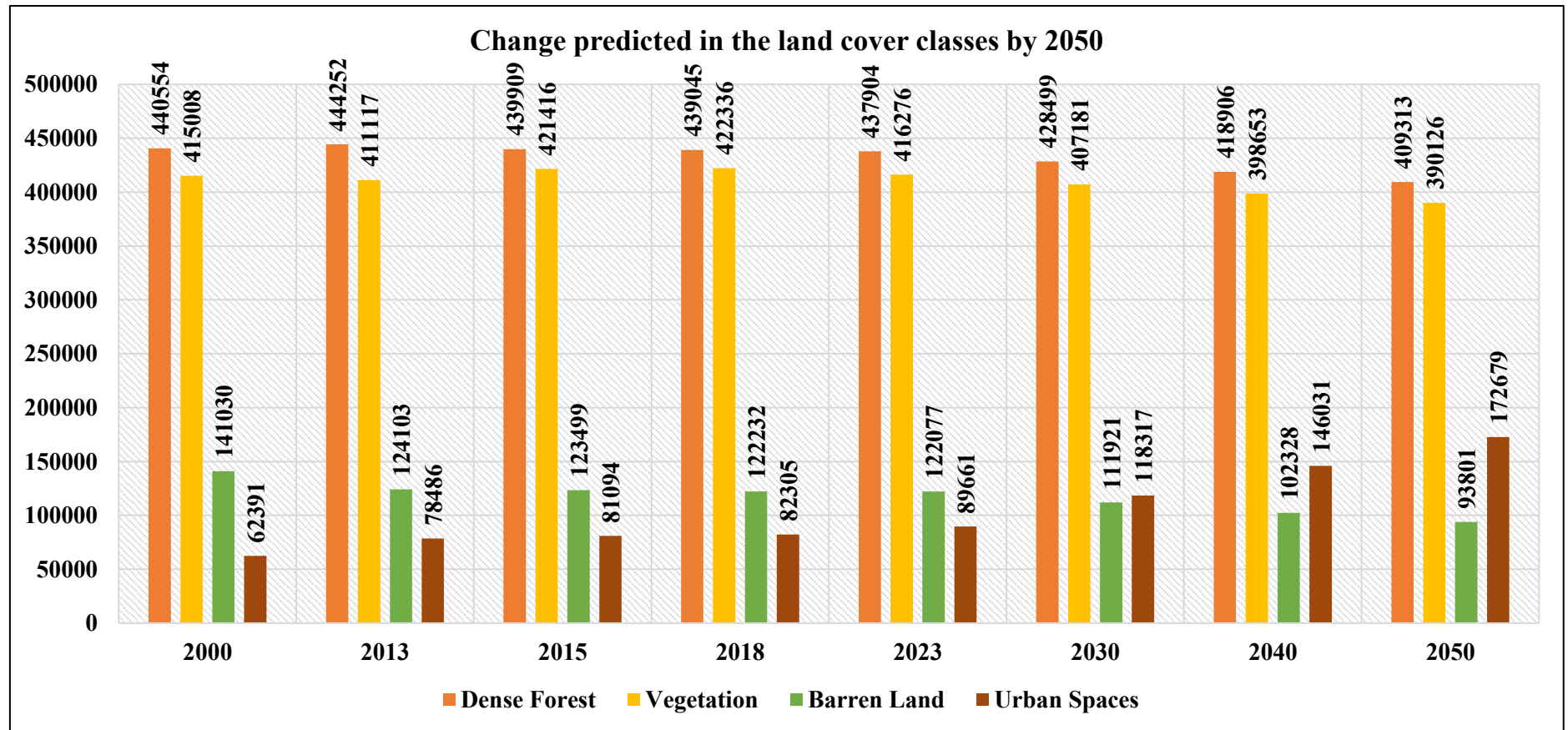


Figure 6.2: Change predicted in the land cover classes by 2050

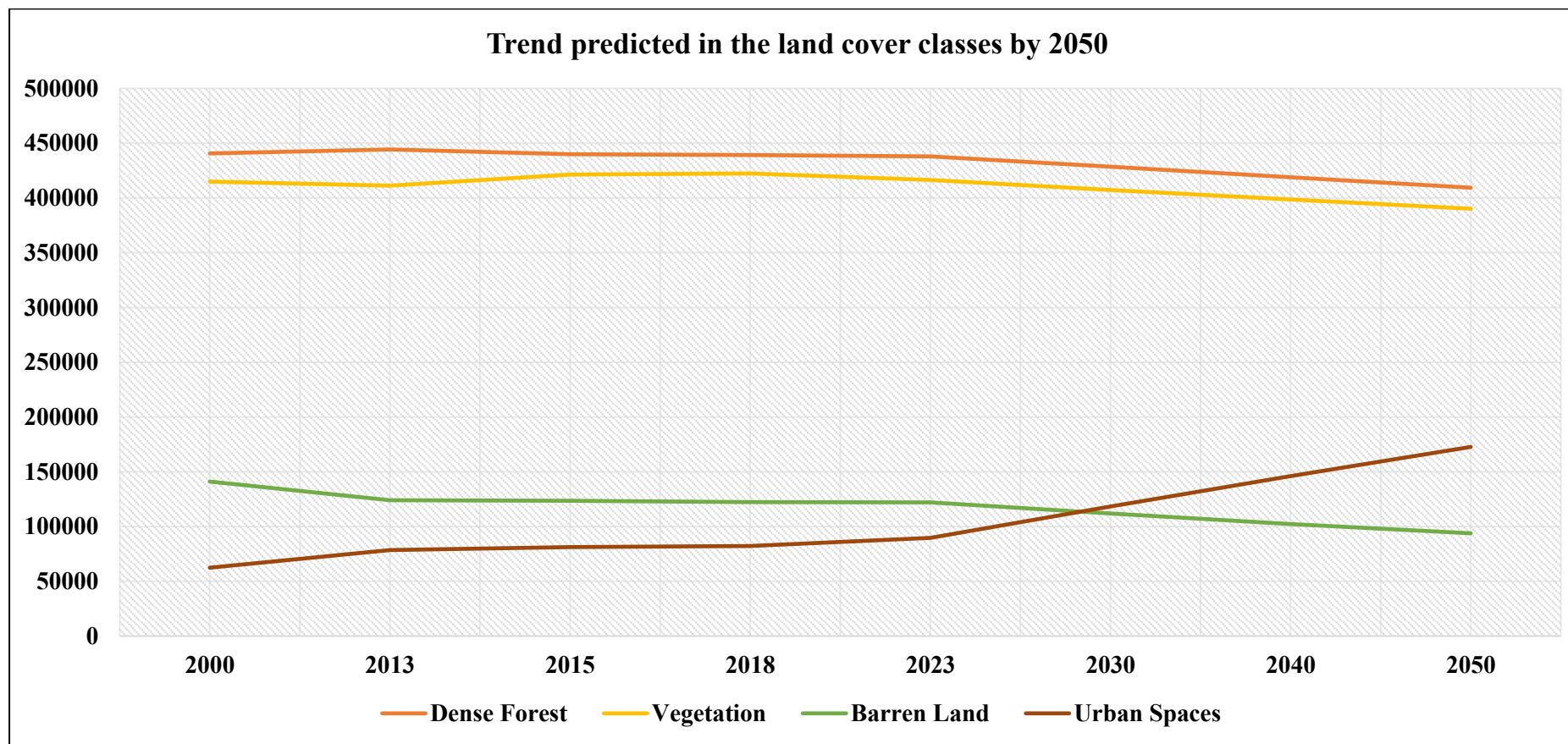


Figure 6.3: Trend predicted in the land cover classes by 2050

6.6.2 Prediction of changes in the cluster density around the 8 seed points

Result obtained by applying algorithm 6.1 to the density of the clusters obtained through Seed-based clustering technique presented in table 4.2 is represented in table 6.2

Table 6.2: Prediction of cluster density around the 8 seed points by 2050

Seed Point	2000	2013	2015	2018	2023	2030	2040	2050	% increase
Gangtok	1539	3009	3418	3599	3472	<i>3608</i>	<i>3788</i>	3968	157.83
Pakyong	1816	2015	1986	1953	1992	<i>1986</i>	<i>1996</i>	2006	10.46
Lingtam	1048	1174	1996	1410	1380	<i>1324</i>	<i>1224</i>	1124	7.25
Rhenock	928	793	690	975	974	<i>1011</i>	<i>1051</i>	1091	17.56
Singtam	1342	1587	1337	1453	1455	<i>1483</i>	<i>1513</i>	1543	14.98
Tumin	284	563	677	732	871	<i>1037</i>	<i>1207</i>	1377	384.86
Sajong	1513	1225	1153	1218	1333	<i>1383</i>	<i>1453</i>	1523	0.66
Ranipool	1067	1790	1935	1827	1833	<i>1893</i>	<i>1983</i>	2073	94.28

Inference 6.2: Trend predicted in the cluster density around the 8 seed points by 2050 (the predicted values and the trend are presented in figure 6.4 and figure 6.5)

By 2050,

- Gangtok, Tumin, and Ranipool is expected to experience an unprecedented increase of 157.83 %, 384.86 % and 94.28 % respectively.
- Pakyong, Lingtam, Rhenock, Singtam, and Sajong is expected to experience a marginal increase 10.46 %, 7.25 %, 17.56 %, 14.98 % and 0.66 % respectively.

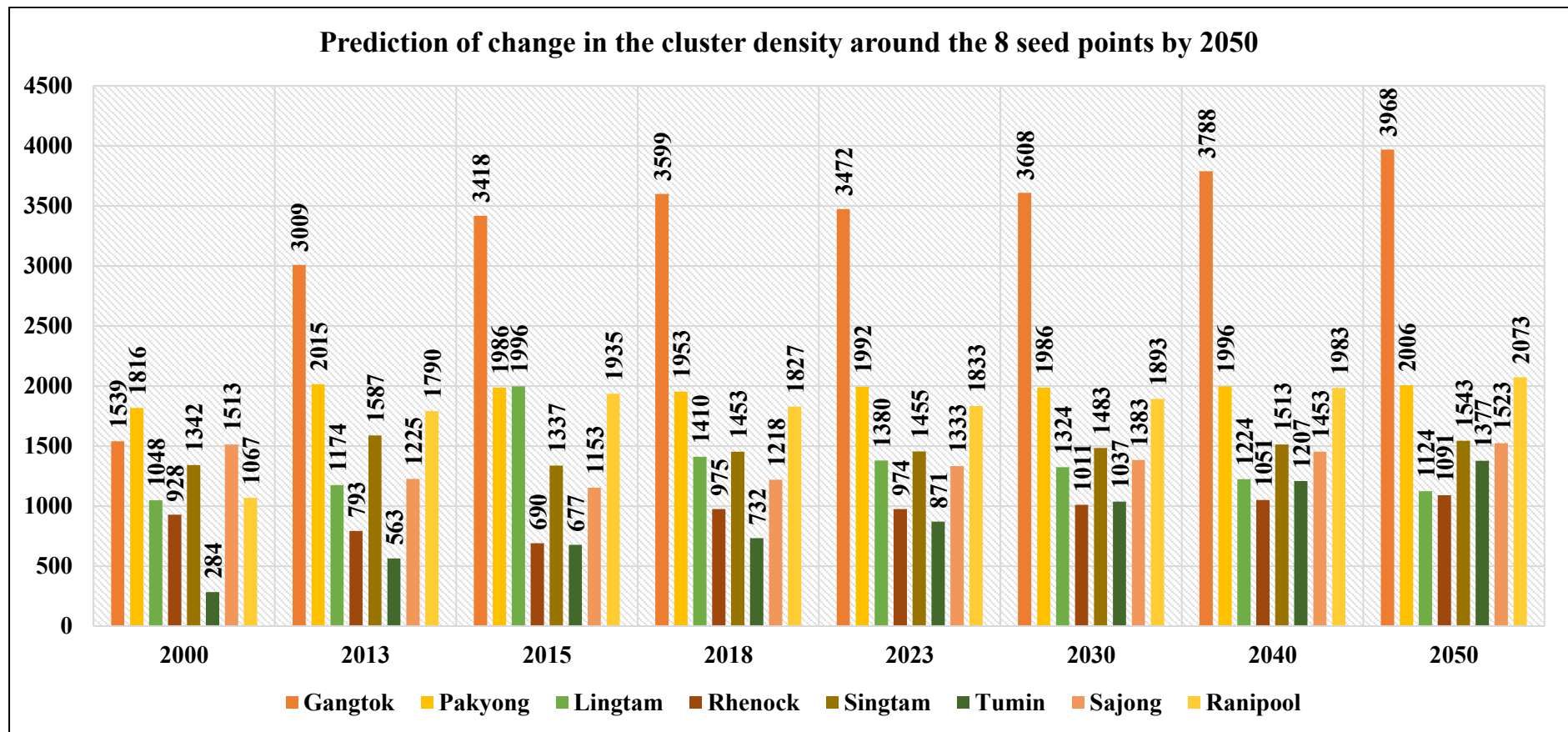


Figure 6.4: Prediction of change in the cluster density around the 8 seed points by 2050

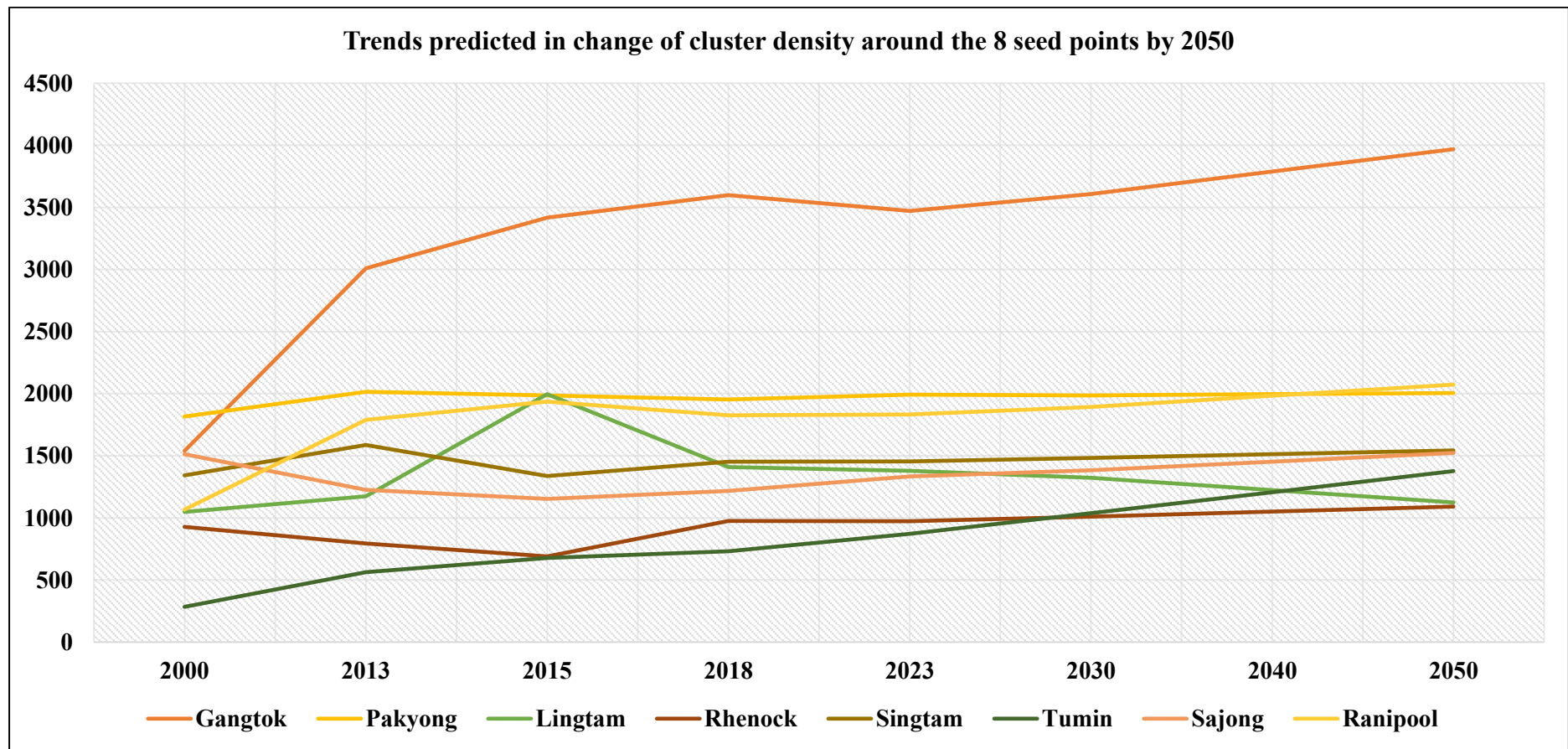


Figure 6.5: Trends predicted in change of cluster density around the 8 seed points by 2050

6.6.3 Prediction of change in the distribution of population participation

Result obtained by applying algorithm 6.1 to the distribution of population presented in 5.1 is represented in table 6.3.

Table 6.3: Prediction of change in the distribution of population participation by 2052

Description	1981	1991	2001	2011	2022	2032	2042	2052	% increase
Urban Sikkim	51084	36984	59870	153578	164432	196886	230554	264221	417.23
Urban East Sikkim	43242	31821	52852	122487	131201	156833	182974	209115	383.59
Urban Rest of Sikkim	7842	5163	7018	31091	33231	40052	47579	55106	602.70
Rural Sikkim	265301	368521	480981	456999	473618	560469	609620	658771	148.31
Total Population	316385	405505	540851	610577	638050	757355	840174	922993	191.73

Inference 6.3: Trend predicted in the distribution of population participation by 2052 (the predicted values and the trend are presented in figure 6.6 and figure 6.7)

- The urban population of the state is predicted to reach 264222 by 2052 of which 209115 is predicted to be contributed by East Sikkim District accounting of 79.14 %. The contribution of all other districts to the urban is predicted at 55107 accounting for only 21 %.
- The predicted percentage increase in the urban population of the entire state is estimated at 417.23 %, likewise for the East Sikkim District it is predicted at 383.59 % and for rest of the districts it is predicted at 602.71 % with an average share of 200.9 % per district.

- The rural population of the state is predicted to reach 658771 by 2052 with a percentage increase of 148.31 %. The total population of the state is predicted to reach 922993 by 2052 with a percentage increase of 191.73 %.
- Therefore, it can be inferred that the predicted percentage increase in the urban population of the East District is considerable higher compared to the other proportions.

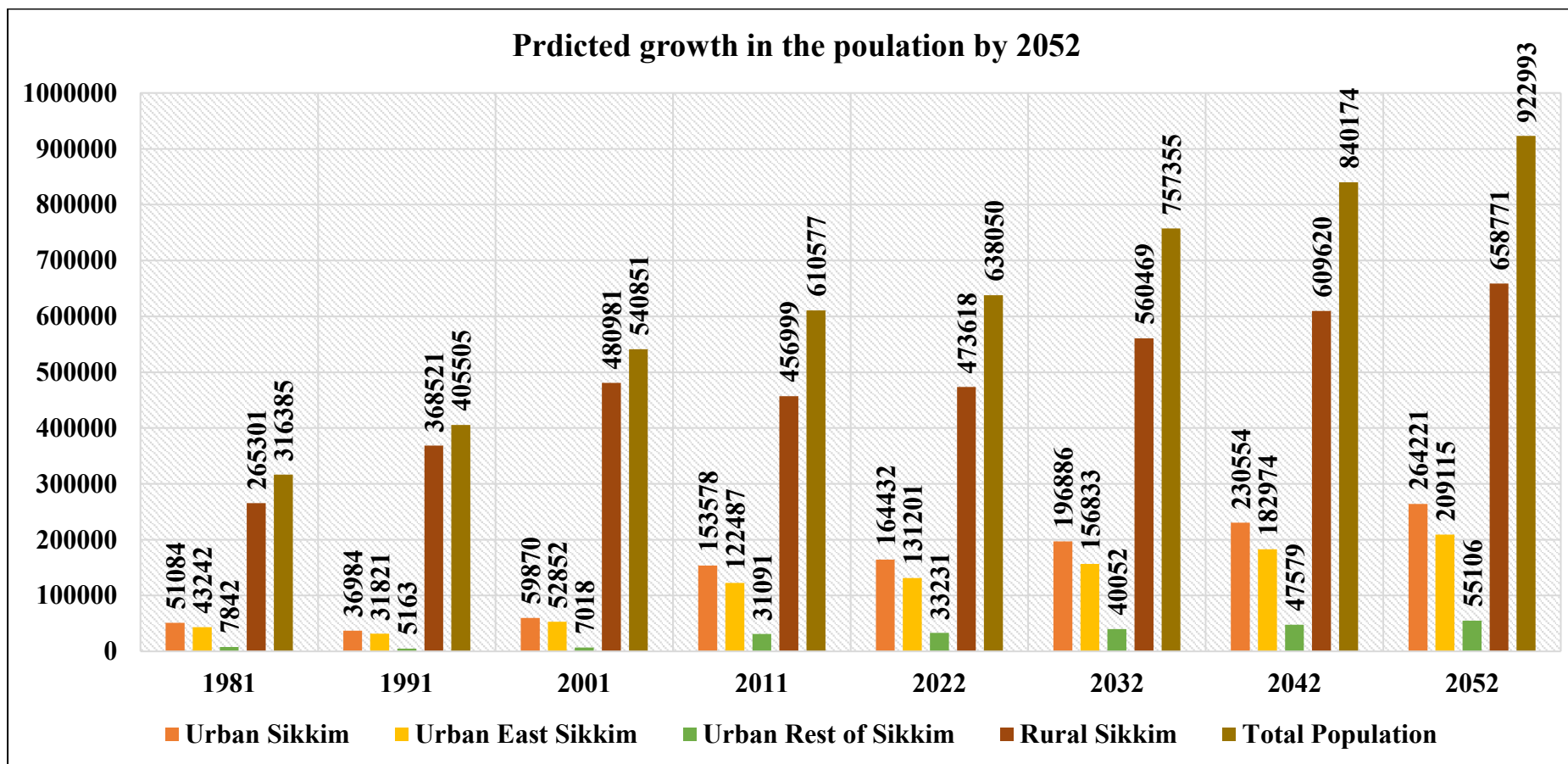


Figure 6.6: Predicted growth in the population by 2052

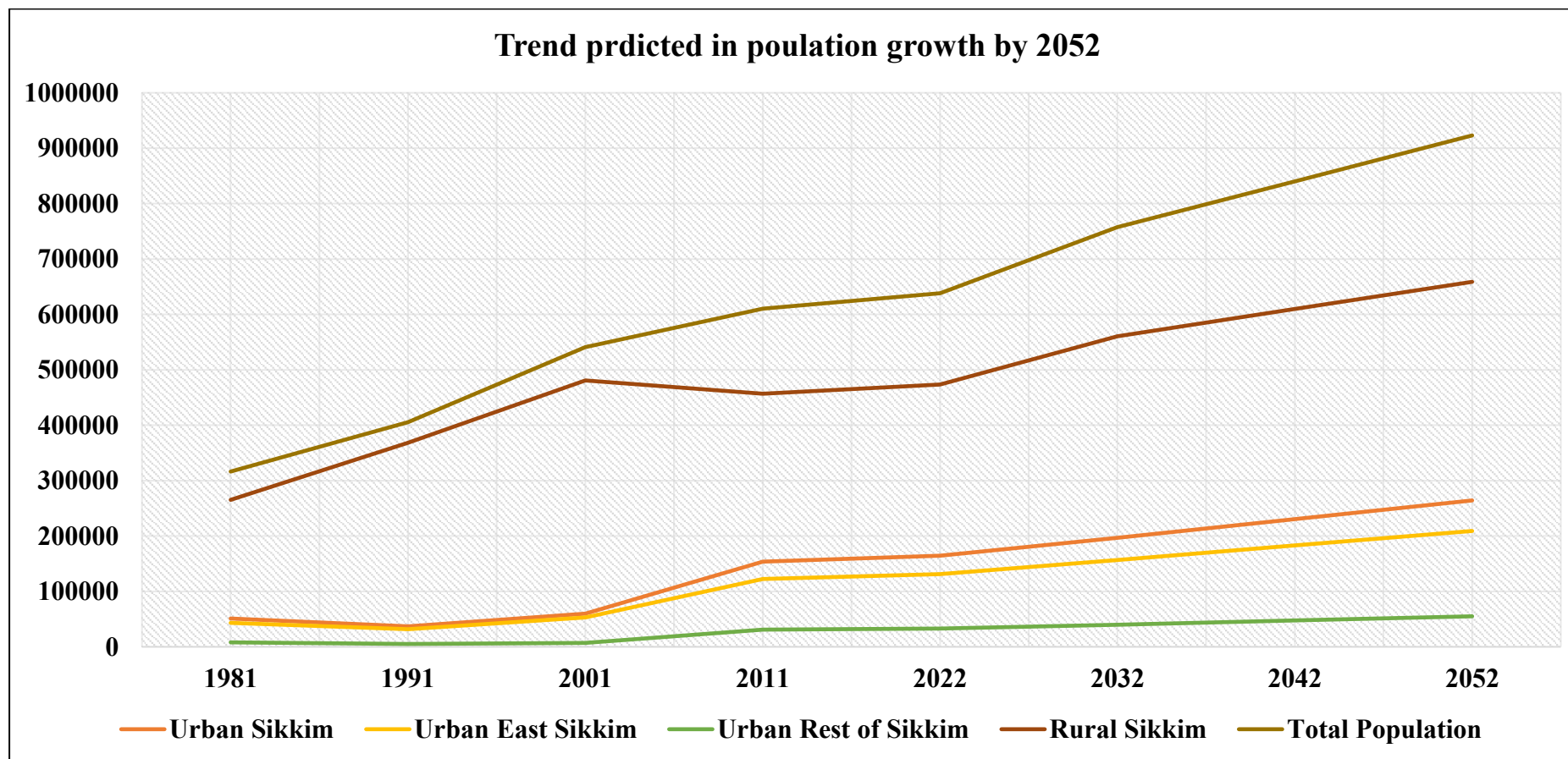


Figure 6.7: Trends predicted in the population growth by 2052

6.6.4 Prediction of change in the household distribution

Result obtained by applying algorithm 6.1 to the household distribution presented in 5.4 is represented in table 6.4.

Table 6.4: Prediction of change in household distribution by 2052

Household	1981	1991	2001	2011	2022	2030	2040	2050	% increase
Urban East Sikkim	9063	10334	11605	28691	39719	<i>42476</i>	<i>50321</i>	58166	541.80
Urban Rest of Sikkim	1603	1498	1393	7027	8985	<i>9851</i>	<i>11848</i>	13844	763.63
Urban Sikkim	10666	11832	12998	35718	48704	<i>52328</i>	<i>62170</i>	72011	575.15
Rural East Sikkim	18218	30597	42976	32876	41364	<i>46872</i>	<i>51617</i>	56362	209.38
Rural Rest of Sikkim	30973	44611	58249	60412	72005	<i>80849</i>	<i>90432</i>	100015	222.91
Rural Sikkim	49191	75208	101225	93288	113369	<i>127721</i>	<i>142049</i>	156377	217.90
Total East Sikkim	27281	40931	54581	61567	81083	<i>89348</i>	<i>101938</i>	114529	319.81
Total Rest of Sikkim	32576	46109	59642	67439	80990	<i>90700</i>	<i>102280</i>	113860	249.52
Total Sikkim	59857	87040	114223	129006	162073	<i>180049</i>	<i>204219</i>	228388	281.56

Inference 6.4: Trend predicted in the distribution of number of households (the predicted values and the trend are presented in figure 6.8 and figure 6.9)

By 2052,

- The total urban household of the state is predicted to reach 72011 by 2052 of which 58166 is predicted to be contributed by East Sikkim District accounting of 80.77 %. The contribution of all other districts of to the urban households is predicted at 13844 accounting for only 19.22 %.
- The predicted percentage increase in the urban household of the entire state is estimated at 575.15 %, likewise for the East Sikkim District it is predicted at 541.80 % and for rest of the districts it is predicted at 763.63 % with an average share of 254.54 % per district.

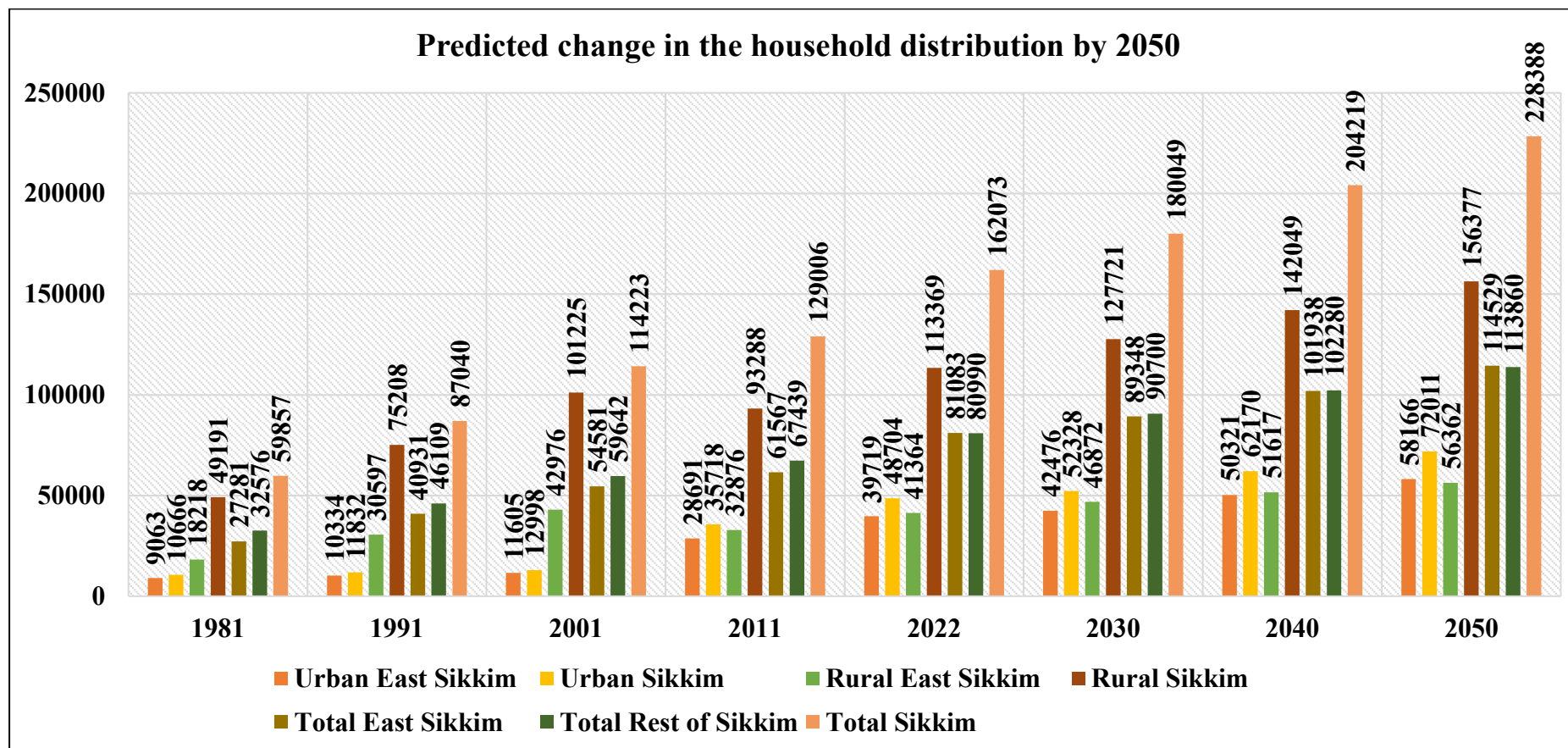


Figure 6.8: Predicted change in the household distribution by 2050

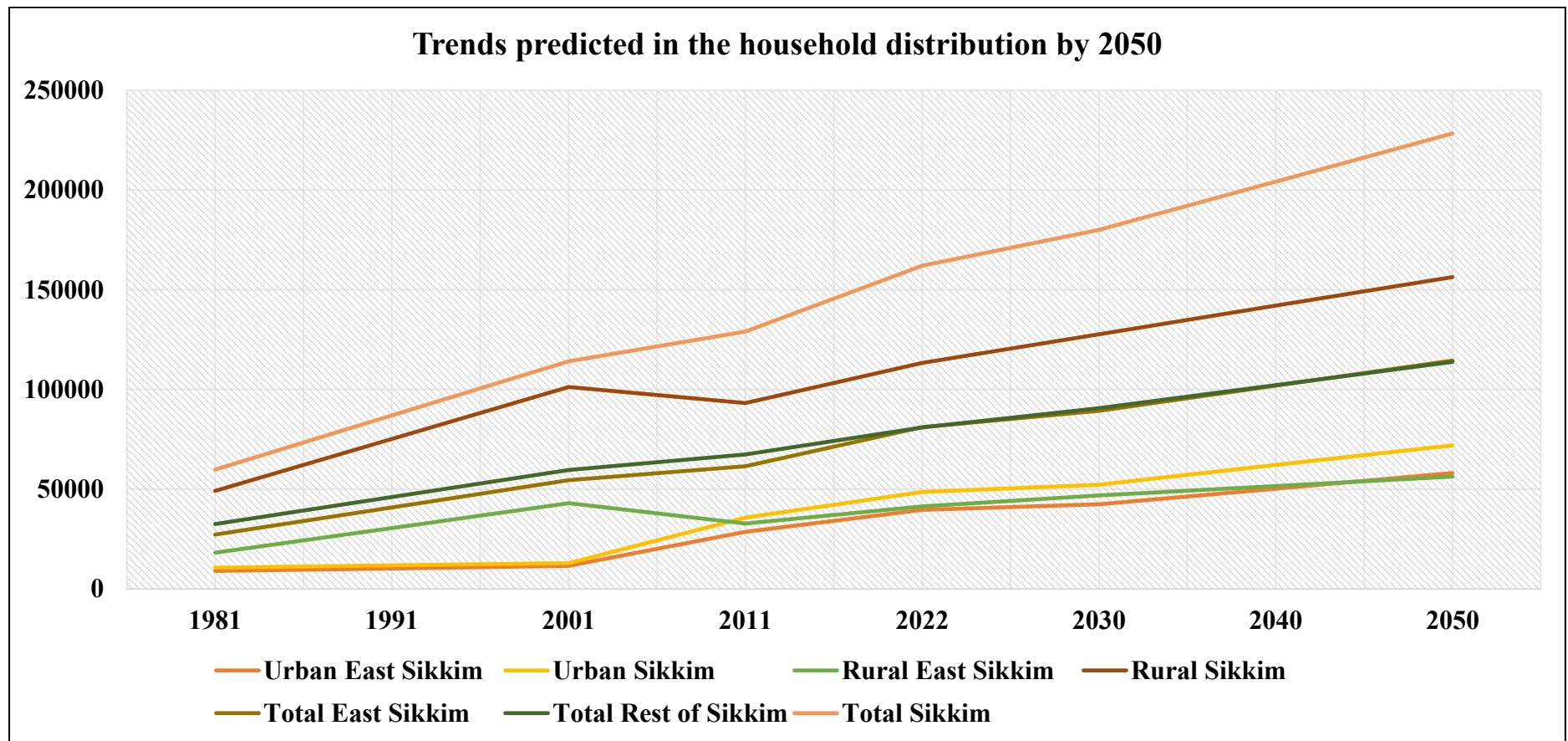


Figure 6.9: Trends predicted change in the household distribution by 2050

6.7 Some Suggestive Measures

i) Promote and incentivize smart growth initiatives.

- Establish clear delineation of zones and promote investment initiatives.
- Implement stringent urban growth boundaries to manage urban expansion.
- Enforce policies and frameworks to ensure sustainable development.
- Advocate for mixed-use development projects
- Support ecologically sustainable practices.
- Develop initiatives centered around transit hubs to decrease dependencies.
- Enhance accessibility to fundamental necessities
- Encourage conservation and sustainable utilization of natural resources.
- Stimulate compact, clustered growth patterns
- Limit development in ecologically vulnerable regions.
- Establish conservation zones and curtail developmental projects.

ii) Strengthen the urban plan framework.

- Develop initial sustainable growth strategies.
- Enforce stringent zoning regulations
- Limit development initiatives to brownfields while conserving greenfields.
- Encourage green infrastructure in urban environments.
- Assign roles and responsibilities to urban residents.
- Construct a comprehensive road network and enhance commuting options to diminish reliance on automobiles.

iii) Rejuvenate existing urban spaces

- Advocate for sustainable infill development initiatives.

- Support the adaptive reuse of present and aging infrastructure.
- Design high-density affordable housing projects
- Enhance public services, including education, transportation, public health, and safety.

iv) Provide convenient and affordable public transport facilities.

- Enhance public transportation infrastructure and encourage its utilization.
- Establish an inexpensive and secure transportation network.
- Construct pedestrian pathways and advocate for their use.
- Promote healthy commuting options, such as cycling.

v) Raise public awareness

- Coordinate awareness activities.
- Inform the public about the advantages of sustainable development.
- Illuminate the expenses associated with urban sprawl.
- Encourage public involvement in the formulation of urban plans.
- Acknowledge sustainable initiatives

vi) Incentivize sustainable growth

- Offers subsidies for sustainable development
- Offers tax benefits for conservation initiatives

6.8 Conclusion

In this section of the research initiative, the results obtained from chapter 3, chapter 4, and chapter 5 were considered for predictive analysis. Linear regression was used for

predicting growth in population, household, and cluster density, whereas Dirichlet regression was used for predicting the growth in land use classes. The selection was based on the criterion of bound; specifically, for land use classes, regardless of their progression, the total number of pixels representing the district will always remain constant. This characteristic is one of the most significant achievements of Dirichlet Regression, as it consistently confines proportion contributions to 1. This technique can also be applied to the results obtained from seed-based clustering when the cluster size becomes unmanageable.

On applying predictive analysis to the land use classes obtained from classification of temporal satellite images, the results reflected that forest cover is expected to decrease by 7.09 %, vegetation is expected to decrease by 0.06 %, barren land is expected to decrease by 0.033 %, and urban sprawl is expected to increase by 1.77 %.

On applying predictive analysis to the outcome obtained from seed-based clustering, the results reflect that Gangtok, Tumin, and Ranipool are expected to experience an unprecedented increase of 157.83 %, 384.86%, and 94.28%, respectively, and Pakyong, Lingtam, Rhenock, Singtam, and Sajong are expected to experience a marginal increase of 10.46 %, 7.25 %, 17.56 %, 14.98%, and 0.66%, respectively.

On applying predictive analysis to the outcome of population estimation, the urban population of the state is predicted to reach 264,222 by 2052, of which 209,115 is predicted to be contributed by East Sikkim District, accounting for 79.14%. The contribution of all other districts to the urban population is predicted at 55,107, accounting for only 21 %. The predicted percentage increase in the urban population

of the entire state is estimated at 417.23%; likewise, for the East Sikkim District, it is predicted at 383.59%, and for the rest of the districts, it is predicted at 602.71%, with an average share of 200.9 % per district. Therefore, it can be inferred that the percentage increase in the urban population of the East District is considerably higher compared to the other.

On applying predictive analysis to the outcome obtained from household estimation, the total urban household of the state is predicted to reach 72,011 by 2052, of which 58,166 is predicted to be contributed by East Sikkim District, accounting for 80.77%. The contribution of all other districts of the urban is predicted at 13844, accounting for only 19.22 %. The predicted percentage increase in the urban household of the entire state is estimated at 575.15%; likewise, for the East Sikkim District, it is predicted at 541.80%, and for the rest of the districts, it is predicted at 763.63%, with an average share of 254.54 % per district.

Thereafter, section also highlights some suggestive measures for effectively containing the influence of urban sprawl.